Accurate modeling of region data

Guido Proietti[1] Christos Faloutsos[2]

April 1998

CMU-CS-98-126

School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

[1]On leave from Dipartimento di Matematica Pura ed Applicata, University of L'Aquila, Via Vetoio, I-67010, Italy. His research was partially supported by the Italian National Research Council (CNR) under the fellowship N.215.29 and by the EU TMR Grant CHOROCHRONOS. E-mail: proietti@cs.cmu.edu.

[2]The research of this author was partially supported by the NSF under Grant IRI-9625428, by the CMU's InforMedia, by the NSF, ARPA and NASA under NSF Cooperative Agreement No. IRI-9411299. E-mail: christos@cs.cmu.edu

Keywords: Spatial databases, GIS, region data, range queries, selectivity estimation, fractals.

Abstract

Spatial data appear in numerous applications, such as GIS [8, 9, 18], multimedia [6] and even traditional databases. Most of the analysis has focused on point data, typically using the uniformity assumption, or, more accurately, a fractal distribution [5]. However, no results exist for non-point spatial data, like 2-d regions (e.g., islands), 3-d volumes (e.g., physical objects in the real world) etc.

This is exactly the problem we solve in this paper. Based on experimental evidence that real areas and volumes follow a "power law", that we named *REGAL* (REGion Area Law), we show (a) the theoretical implications of our model and its connection with the ubiquitous fractals and (b) the first of its practical uses, namely the selectivity estimation for range queries. Experiments on a variety of real datasets (islands, lakes, human-inhabited areas) show that our method is extremely accurate, enjoying a maximum relative error ranging from 1 to 5%, versus 30-70% of a naive model that uses the uniformity assumption.

1 Introduction

Spatial data appear in numerous applications, such as GIS [8, 9, 18], multimedia [6] and spatiotemporal databases [19]. Statistical modeling of real data involves the concise description of a dataset with a few parameters (e.g., total count, area, etc.), so that we can obtain accurate estimates. Such a concise description is useful for at least three settings:

- selectivity for range queries, k nearest neighbor queries, spatial joins etc.
- analysis of spatial access methods (SAM). For example, how many nodes will an R-tree or quadtree require to store the data, how many such nodes a query will touch, etc.
- generation of pseudo-random, but realistic, spatial datasets, that can stress-test SAMs, whenever real data are not available. For example, for scale-up studies, or for studies on high dimensionalities [2], where we want to control the statistics, to do, e.g., sensitivity analysis.

Most of the analysis efforts have focused on point data, typically using the uniformity assumption, or, more accurately, a fractal distribution [5]. In fact, for point data, these two numbers (the count and the fractal dimension of a dataset) are sufficient to accurately estimate selectivities for range queries, spatial joins and nearest neighbor queries, as we describe in the survey section.

However, no results exist for non-point spatial data, like 2-d regions (islands, lakes, vegetation patches) and 3-d volumes (physical objects in the real world). We shall refer to such data that have non-zero d-dimensional volume as *region data*, although the upcoming discussion holds for any dimensionality of the address space. Thus, the problem we focus on is the following: *We are given a real set of region data (normalized on the unit d-dimensional hypercube); what is the smallest number of parameters that we need to describe it?*

For point data, the count and the fractal dimension are enough. However, for region data, it is not even clear what we mean by fractal dimension: d-dimensional region objects have fractal dimension equal to d. Maybe we want the fractal dimension of the centers of our region data? Or maybe something else?

We answer all these questions in the rest of the paper, developing a realistic statistical model for region data, and showing how to use it to compute the selectivity of range queries. Its maximum relative error ranges from 1 to 5%, versus 30-70% of the naive model, based on the uniformity assumption.

The paper is organized as follows: Section 2 gives a brief description of previous work on the topic. In Section 3 we develop our model and we show how it can be used to estimate selectivity of window queries on regions datasets in the d-dimensional space. Section 4 provides a large collection of experimental results on real region data (collection of islands, lakes, urban areas, etc.). Section 5 presents the relationship existing between our model and the fractal theory, exploit it to provide a realistic random region generator and suggest some directions for a practitioner for an effective application of our model. Finally, Section 6 contains concluding remarks and future work.

2 Survey

The main topic within the spatial database field which is related to our present work is *query optimization*, and, more specifically, *selectivity estimation* of range (or window) queries, which are the most popular spatial access operation [15, 13]. In the database community, there is a mounting evidence that query optimization is becoming more and more important with the advent of spatial databases consisting of petabytes of data and, in the near future, of huge spatiotemporal databases [4].

In [11, 15], an analytical formula to compute selectivity for a window query as a function of the underlying data morphology and distribution is given. To apply such a formula when these parameters are unknown, one typically makes the *uniformity* and *independence* assumption on them. Unfortunately, these assumptions do not hold for real datasets and generally lead to pessimistic results [3]. Recently, the introduction of the concept of fractal dimension has allowed to better describe the statistical properties of the data themselves and to precisely analyze space and time performances of spatial data structures used to store them. Using the fractal dimension, we can accurately estimate the performance of R-tree for range queries [5], the selectivity of spatial joins [1] and the performance of nearest-neighbor queries [16]. However, all these works focus on *point* data only. Therefore, to the best of our knowledge, this is the first attempt to model accurately region datasets.

3 Proposed method

In this section, we first give the problem definition and we show a naive solution. After, we give the proposed solution, for $d = 2$ dimensions and for arbitrary d. Table 1 gives a list of symbols used throughout this section.

3.1 Problem definition

Let us rigorously state the problem we are concerned with. For the sake of clarity, let us first focus on the 2-dimensional space. Later, all the results will be extended to the d-dimensional space.

PROBLEM: selectivity of rectangles
Given:

- A set of *similar* rectangles (i.e., having a fixed given aspect ratio ρ between width and height) $\mathcal{R} = \{r_1, r_2, \ldots, r_N\}$ embedded in $U = [0,1] \times [0,1]$.

Symbol	Definition
$\mathcal{R}$	Dataset of rectangles
N	Total number of rectangles of $\mathcal{R}$
A	Total area of $\mathcal{R}$
W	Total width of $\mathcal{R}$
H	Total height of $\mathcal{R}$
ρ	Fixed ratio between width and height of rectangles in $\mathcal{R}$
$a_{\max}$	Area of the largest rectangle in $\mathcal{R}$
$w_{\max}$	Width of the biggest rectangle in $\mathcal{R}$
$h_{\max}$	Height of the largest rectangle in $\mathcal{R}$
B	Patchiness exponent
$C(a)$	Number of regions having area at least a
$C_W(w)$	Number of regions having width at least w
$C_H(h)$	Number of regions having height at least h
$\vec{q} = (q_1, \ldots, q_d)$	Query window of sides $q_1, \ldots, q_d$
$Sel(\mathcal{R}, \vec{q})$	Avg. selectivity for range queries of sides $q_1, \ldots, q_d$
U	Image space

Table 1: Symbol table

- The total area A of $\mathcal{R}$.
- A $q_x \times q_y$ *window query* $\vec{q}$.

Find the *selectivity* $Sel(\mathcal{R}, \vec{q})$ in $\mathcal{R}$ of the window query $\vec{q}$, that is, the number of rectangles in $\mathcal{R}$ intersecting $\vec{q}$.

The formula in [11, 15] gives the selectivity when we know the width w_i and the height h_i of every rectangle. Hence

$$Sel(\mathcal{R}, \vec{q}) = \sum_{i=1}^{N} (q_x + w_i)(q_y + h_i), \tag{1}$$

which can also be written

$$Sel(\mathcal{R}, \vec{q}) = A + q_x \cdot H + q_y \cdot W + q_x \cdot q_y \cdot N \tag{2}$$

where W and H are the total width and height extent of $\mathcal{R}$. The question is to estimate the selectivity with much less information.

3.2 Naive solution

The major problem is the following: what assumption should we make about rectangle sizes distribution? Is it Gaussian? Is it bimodal? Clearly, the most straightforward assumption is to assume that sizes are uniform. In this case, being rectangles similar with aspect ratio ρ, we can merely conclude that the expected width of a rectangle in $\mathcal{R}$ is $\sqrt{\rho \cdot \frac{A}{N}}$, while its expected height is $\sqrt{\frac{1}{\rho} \cdot \frac{A}{N}}$. Therefore, since

$$W = \sqrt{\rho \cdot \frac{A}{N}} \cdot N \quad \text{and} \quad H = \sqrt{\frac{1}{\rho} \cdot \frac{A}{N}} \cdot N$$

it follows from (2) that the expected selectivity is

$$Sel(\mathcal{R}, \vec{q}) = A + \left(q_x \cdot \sqrt{\frac{1}{\rho}} + q_y \cdot \sqrt{\rho} \right) \cdot \sqrt{A \cdot N} + q_x \cdot q_y \cdot N. \tag{3}$$

3.3 Proposed solution: the REGAL law

However, real region datasets do not obey the uniformity assumption. Rather, it turns out that the complementary cumulative distribution function[1] (CCDF) of the areas of the regions obeys the following hyperbolic power law:

REGion Area Law (REGAL): *The number of regions $C(a)$ of area greater than or equal to a follows the hyperbolic power law*

$$\boxed{C(a) = k \cdot a^{-B} \quad k, B > 0, \quad a \geq 0.} \tag{4}$$

Korčak was the first to observe such a law, for the Aegean Islands (he suggested $B \approx 0.5$) [12]. The exponent B is also called the *patchiness exponent*. Recent measurements on 2-d region datasets from diverse applications suggest that usually a similar power law holds [10], with B in the range $[0.5, 0.9]$. In Section 5, we show that the power law (4) is related to fractals. Given that fractals appear surprisingly often in nature, we expect that the majority of real region datasets will obey (4). Moreover, as a consequence of the inherent self-similarity of real region datasets, the minimum bounding rectangles (MBRs) of the regions are expected to follow the same law as well.

Under the realistic assumption that rectangles in $\mathcal{R}$ obey to (4), we now show that we can compute much more accurate estimates on the selectivity if we are given the patchiness exponent B. Notice that the uniformity assumption is unable to use this extra information. We prove the following:

[1]Remember that the cumulative distribution function of $f(x) : \Re \rightarrow \Re$ is defined as $F(x) = \int_{-\infty}^{x} f(t)dt$, while the complementary cumulative distribution function is defined as $\overline{F}(x) = \int_{x}^{+\infty} f(t)dt$.

Theorem 1 *Given a set $\mathcal{R} = \{r_1, r_2, \ldots, r_N\}$ of rectangles embedded in U whose areas obey to the REGAL law, having a fixed given aspect ratio ρ between width and height, a total area A and a patchiness exponent B, the selectivity of a rectangular window query $\vec{q}$ is*

$$\boxed{Sel(\mathcal{R}, \vec{q}) = A + \left(q_x \cdot \sqrt{\tfrac{1}{\rho}} + q_y \cdot \sqrt{\rho}\right) \cdot \sqrt{A \cdot \frac{1-\frac{1}{B}}{N^{\left(1-\frac{1}{B}\right)}-1}} \cdot \left(\frac{N^{\left(1-\frac{1}{2B}\right)}-1}{1-\frac{1}{2B}}\right) + q_x \cdot q_y \cdot N.} \tag{5}$$

Proof. We start with (2). We need to estimate the sum of widths W and the sum of heights H. By assumption, $C(a)$ obeys to the REGAL law (4). Hence, from the initial condition $1 \equiv C(a_{\max}) = k \cdot a_{\max}^{-B}$, where $a_{\max}$ is the (unknown) area of the largest rectangle, it follows that

$$C(a) = a_{\max}^{B} \cdot a^{-B}.$$

We need to estimate $a_{\max}$. From the inverse relation, we have

$$a(C) = \left(\frac{1}{a_{\max}^{B}} \cdot C\right)^{-\frac{1}{B}} = a_{\max} \cdot C^{-\frac{1}{B}}$$

Therefore, if a_i denotes the area of the i-th rectangle of $\mathcal{R}$, it follows

$$A = \sum_{i=1}^{N} a_i \approx a_{\max} \int_1^N C^{-\frac{1}{B}}\, dC = a_{\max} \cdot \left[\frac{C^{\left(1-\frac{1}{B}\right)}}{1-\frac{1}{B}}\right]_1^N = a_{\max} \cdot \frac{N^{\left(1-\frac{1}{B}\right)}-1}{1-\frac{1}{B}}$$

from which it follows

$$a_{\max} = A \cdot \frac{1-\frac{1}{B}}{N^{\left(1-\frac{1}{B}\right)}-1}. \tag{6}$$

Let now $C_W(w)$ and $C_H(h)$ denote the number of regions in $\mathcal{R}$ having width and height at least w and h, respectively. Since the rectangles in $\mathcal{R}$ are similar, we have $a = \frac{1}{\rho}w^2$ and $a = \rho h^2$. Then, from (4), $C_W(w)$ obeys the following power law

$$C_W(w) = (\rho \cdot a_{\max})^B \cdot w^{-2B} \tag{7}$$

and analogously for $C_H(h)$

$$C_H(h) = \left(\frac{a_{\max}}{\rho}\right)^B \cdot h^{-2B}. \tag{8}$$

Denoting with

$$w_{\max} = \sqrt{\rho \cdot a_{\max}} = \sqrt{\rho \cdot A \cdot \frac{1-\frac{1}{B}}{N^{\left(1-\frac{1}{B}\right)}-1}} \tag{9}$$

the width of the largest rectangle, and with

$$h_{\max} = \sqrt{\frac{a_{\max}}{\rho}} = \sqrt{\frac{1}{\rho} \cdot A \cdot \frac{1-\frac{1}{B}}{N^{\left(1-\frac{1}{B}\right)}-1}} \tag{10}$$

its height, (7, 8) can be written

$$C_W(w) = w_{\max}^{2B} \cdot w^{-2B}$$

and

$$C_H(h) = h_{\max}^{2B} \cdot h^{-2B}.$$

Hence, from the inverse relations we have

$$w(C_W) = \left(\frac{1}{w_{\max}^{2B}} \cdot C_W\right)^{-\frac{1}{2B}} = w_{\max} \cdot C_W^{-\frac{1}{2B}}$$

and

$$h(C_H) = \left(\frac{1}{h_{\max}^{2B}} \cdot C_H\right)^{-\frac{1}{2B}} = h_{\max} \cdot C_H^{-\frac{1}{2B}}.$$

Therefore, if w_i (h_i) denotes the width (height) of the i-th rectangle of $\mathcal{R}$, it follows

$$W = \sum_{i=1}^{N} w_i \approx w_{\max} \int_1^N C_W^{-\frac{1}{2B}}\ dC_W = w_{\max} \cdot \left[\frac{C_W^{\left(1-\frac{1}{2B}\right)}}{1-\frac{1}{2B}}\right]_1^N = w_{\max} \cdot \frac{N^{\left(1-\frac{1}{2B}\right)}-1}{1-\frac{1}{2B}} \tag{11}$$

and similarly

$$H = \sum_{i=1}^{N} h_i \approx h_{\max} \int_1^N C_H^{-\frac{1}{2B}}\ dC_H = h_{\max} \cdot \frac{N^{\left(1-\frac{1}{2B}\right)}-1}{1-\frac{1}{2B}}. \tag{12}$$

Therefore, from (2, 9, 10, 11, 12) the thesis follows. □

Observation 1: Notice that for $B = 1$, by applying De l'Hospital's rule, (6) still holds, and more precisely we have:

$$a_{\max} = \lim_{B\to 1} A \cdot \frac{1-\frac{1}{B}}{N^{\left(1-\frac{1}{B}\right)}-1} = \frac{A}{\ln N}.$$

Similarly, for $B = 1/2$, (11, 12) still hold, and we have:

$$W = \lim_{B\to 1/2} w_{\max} \cdot \frac{N^{\left(1-\frac{1}{2B}\right)}-1}{1-\frac{1}{2B}} = w_{\max} \cdot \ln N$$

and analogously, $H = h_{\max} \cdot \ln N$.

Observation 2: Equation (6) establishes a relationship between $a_{\max}$ and A, once N is given. This means that we can provide an accurate estimation of $Sel(\mathcal{R}, \vec{q})$ even if $a_{\max}$ is given instead of A. Notice that in this scenario, using the uniformity assumption, we can merely conclude that $A = a_{\max} \cdot N$, which most likely will heavily overestimate $Sel(\mathcal{R}, \vec{q})$.

As we show next, the above theorem will provide a good estimation for window selectivity on real region datasets.

3.4 Generalization to the d-dimensional space

The above result can be extended to the d-dimensional space. In such a case, the problem is as follows: we are given a set $\mathcal{R} = \{r_1, r_2, \ldots, r_N\}$ of d-dimensional similar (i.e., having a fixed given aspect ratio $\rho_{i,j}$ between the i-th and the j-th side) hyper-rectangles embedded in $U = [0,1]^d$, their total volume V and their patchiness exponent B. Let the CCDF of the volumes of the hyper-rectangles follows the power law

$$C(v) = k \cdot v^{-B}$$

Then, the selectivity of a rectangular window query $\vec{q} = (q_1, \ldots, q_d)$ is

$$Sel(\mathcal{R}, \vec{q}) = V + \sum_{\{i_1,\ldots,i_j\} \in 2^{\{1,\ldots,d\}} \setminus \{\phi, \{1,\ldots,d\}\}} q_{i_1} \cdot \ldots \cdot q_{i_j} \cdot x_{i_{j+1}} \cdot \ldots \cdot x_{i_d} \cdot \left(\frac{N^{\left(1-\frac{d-j}{d \cdot B}\right)} - 1}{1 - \frac{d-j}{d \cdot B}} \right) + \prod_{i=1}^{d} q_i \cdot N$$

where $2^{\{1,\ldots,d\}}$ denotes the power set[2] of $\{1, \ldots, d\}$ and

$$x_i = \sqrt[d]{v_{\max} \cdot \prod_{j=1}^{d} \rho_{i,j}} = \sqrt[d]{V \cdot \frac{1 - \frac{1}{B}}{N^{\left(1-\frac{1}{B}\right)} - 1} \cdot \prod_{j=1}^{d} \rho_{i,j}}$$

is the i-th side of the largest object having volume $v_{\max}$.
The above formula can be proved by analogy from the 2-dimensional case and is here omitted.

The above expression can be simplified when working on set of hyper-cubic objects. In such a case, setting $x_{\max}$ the side of the largest region, we have

$$Sel(\mathcal{R}, \vec{q}) = V + \sum_{\{i_1,\ldots,i_j\} \in 2^{\{1,\ldots,d\}} \setminus \{\phi, \{1,\ldots,d\}\}} q_{i_1} \cdot \ldots \cdot q_{i_j} \cdot x_{\max}^{d-j} \cdot \left(\frac{N^{\left(1-\frac{d-j}{d \cdot B}\right)} - 1}{1 - \frac{d-j}{d \cdot B}} \right) + \prod_{i=1}^{d} q_i \cdot N$$

where

[2] Remember that, gives a set S, the power set of S, denoted as 2^S, is defined as the set of all subsets of S, including the empty set and S itself. If S is finite, the cardinality of 2^S is $2^{|S|}$.

$$x_{\max} = \sqrt[d]{v_{\max}} = \sqrt[d]{V \cdot \frac{1-\frac{1}{B}}{N^{\left(1-\frac{1}{B}\right)}-1}}.$$

The above expression, for square window queries of side q, further simplifies to

$$Sel(\mathcal{R}, \vec{q}) = V + \sum_{j=1}^{d-1} \binom{d}{j} \cdot q^j \cdot x_{\max}^{d-j} \cdot \left(\frac{N^{\left(1-\frac{d-j}{d \cdot B}\right)}-1}{1-\frac{d-j}{d \cdot B}} \right) + q^d \cdot N.$$

4 Experiments on real datasets

To assess experimentally the accuracy of our analysis, we have used three different region datasets, that is:

- The Scandinavian Lakes (LAKES), available at `http://mapweb.parc.xerox.com/map/nogrid` (Xerox PARC Map Viewer) and consisting of 810 lakes.
- The Indonesia Archipelago (ISLANDS), available at `http://mapweb.parc.xerox.com/map/nogrid`, and consisting of 470 islands.
- A population density map of Europe (REGIONS). This map has been created starting from a population density map from a World Atlas. Each grid cell is turned to black if it has density above a threshold, namely 30 inhabitants/Km^2. It consists of 757 regions.

We also used three additional datasets: the Aegean Islands (51 islands) and the Japan Archipelago (186 islands), both available at `http://mapweb.parc.xerox.com/map/nogrid`, and a map of Italy agricultural plains (228 regions), created starting from a geographic map from a World Atlas and turning to black a grid cell whenever it is at most 50 meters above the sea level. We do not give details about these datasets since the results were similar.

In the following subsections we present results for: (a) verifying that the MBRs of the regions obey to the REGAL law (4); (b) verifying the accuracy of our formula (5) as compared to the formula derived using a uniformity assumption (3).

4.1 Verifying the REGAL law

All the datasets were stored using 1024×1024 bitmaps, as shown in Figure 1a-c. Preliminary, we have identified all the regions and their MBRs in each dataset. Then, we have computed all the relevant features needed for checking our results. These data are summarized in Table 2. Note that to estimate B, we have computed the CCDF of the MBRs area for each dataset and we have interpolated the plotted

Dataset	N	A	B	ρ
LAKES	810	75,910	0.85	1.13
ISLANDS	470	136,893	0.60	1.98
REGIONS	757	190,526	0.70	0.53

Table 2: Datasets features.

points with a straight line using the classic least-square method. Note also that ρ has been computed by averaging over all the MBRs' aspect ratios.

Figure 1e-f shows in a log-log diagram the obtained results, along with the regression line, whose negated slope is the patchiness exponent B. It is impressive that the MBRs of all three datasets, even if their characteristics are so different, obey almost perfectly to (4).

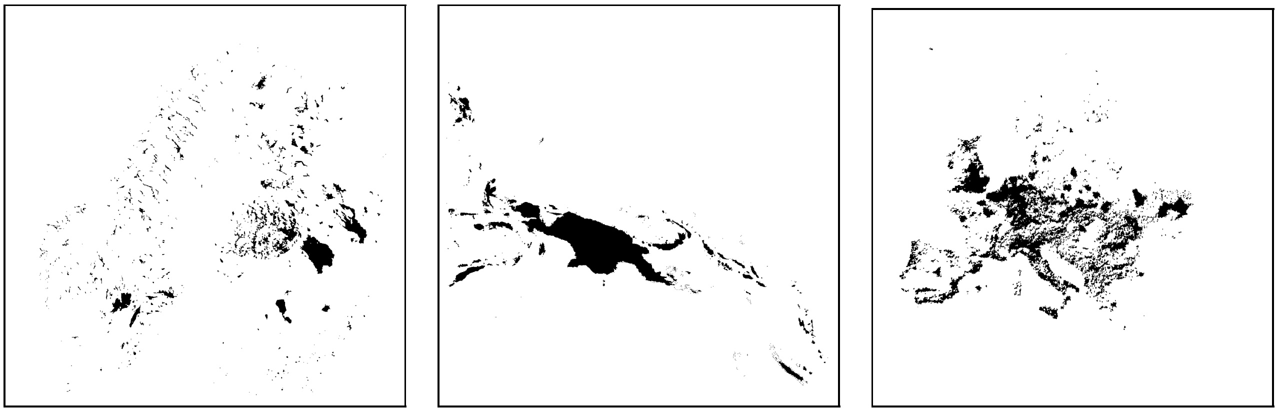
4.2 Accuracy of our selectivity estimation

To ascertain the accuracy of our formula (5) as compared to the formula derived used a uniformity assumption (3), for each dataset we have initially computed the real selectivity[3] using (1). After, we applied (3,5), for query windows of width $2^i, i = 0, \ldots 10$ and having three different aspect ratios: 1:1 (square), 1:2 and 2:1. Figure 2 shows the relative error of our approach, as compared to that of the uniformity model, for the LAKES, ISLANDS and REGIONS dataset, respectively. Note that, for each dataset, our approach is usually within 1% to the reality, and never exceeds a 5% of relative error. On the other hand, the uniformity model can give up to 70% relative error. Note also that the ratio between the relative error of the uniformity model and the proposed model is enormous: in particular, for 1:2 window queries on the REGIONS dataset, it is 44 in the average (i.e., the proposed model is 44 times more precise!). Finally, following the recommendations from statistics, we have also computed the *geometric average* of relative errors, for each dataset and for each different window aspect ratio, summarized in Table 3. Even in this case, the ratio between the two models is huge: in particular, for 1:2 window queries on the REGIONS dataset, it has a peak of 30.

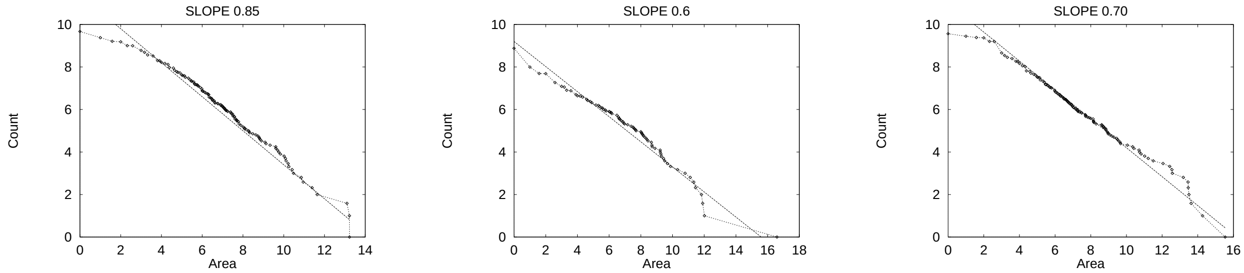
5 Discussion

In this section, we first discuss the relationship existing between the patchiness exponent and the fractal dimension, and we exploit it to provide a realistic random region generator and a fast estimation of B. After, we suggest directions to a practitioner to fully exploit our method application.

[3]Of course, all the computations have been normalized to the 1024×1024 image space.



(a) LAKES (b) ISLANDS (c) REGIONS



(e) LAKES (e) ISLANDS (f) REGIONS

Figure 1: Used datasets: (a) LAKES; (b) ISLANDS; (c) REGIONS, together with their patchiness plots: log(count) *vs* log(area) for (d) LAKES; (e) ISLANDS; (f) REGIONS.

5.1 Fractals, patchiness and random region generators

Power laws go hand-in-hand with self-similarity and fractals [17] and (5) is no exception. Let's see why this is the case, and how we can use fractals to our advantage. First, a quick introduction to fractals is necessary.

Preliminaries: A fractal is a set of points that are exactly or statistically self similar. Exact fractals are generated recursively, by a "generator", applied recursively to an "initiator". Figure 3 gives an example of a famous fractal, the "Koch snowflake". Figure 3a gives a (unit) line segment (the initiator); Figure 3b gives the generator, consisting of $N_r = 4$ smaller segments, each of size $r = 1/3$. Figure 3c shows the replacement of each of the segments of 3b with a miniature replica of the generator. Repeating the process to infinity, we obtain the "Koch curve". Notice that it has infinite length, that is $\lim_{n \to \infty} (4/3)^n$. Gluing 3 such curves together, we obtain the Koch snowflake (Figure 3d).

The *(Hausdorff) fractal dimension* D_H for a strictly self-similar fractal is defined as

Geometric average relative error (%)						
Ratio	1:1		1:2		2:1	
Dataset	REGAL	UNIF	REGAL	UNIF	REGAL	UNIF
LAKES	0.92	8.30	1.08	8.63	0.82	8.43
ISLANDS	0.58	15.66	1.77	14.52	0.69	17.68
REGIONS	0.78	12.63	0.51	15.44	1.85	10.70

Table 3: Geometric average relative error (%) in estimating $Sel(\mathcal{R}, \vec{q})$ of the proposed method (REGAL) as compared to the uniform model (UNIF), for each dataset and for each aspect ratio of the query window.

$$D_H = \frac{\log N_r}{\log(1/r)} \tag{13}$$

and gives a measure of the "roughness" of the fractal. For a straight line, we have $D_H = 1$; for the Koch snowflake, we have $D_H = \log 4/\log 3$, slightly higher than 1, that is, it is more rugged than a straight line. One of the most rugged curves is the Hilbert curve, with fractal dimension $D_H = 2$ [14], hence it is a space-filling curve [7].

σ-fractals: However, fractals like the Koch snowflake consist of a single region. For generating multiple regions, we make use of the so-called $\sigma-$*fractals*. Figure 4 gives a possible $\sigma-$fractal generator, together with the resulting regions after 2 iterations on the sides of the square with side $s_{\max}$.

It turns out that the following theorem holds:

Theorem 2 (Mandelbrot) *For a σ-fractal in a d-dimensional space, we have*

$$B = \frac{D_H}{d}$$

where B is the patchiness exponent of the regions (d-dimensional volumes) and D_H is the fractal dimension of their boundaries.

Proof. See [14]. □

Given the inherent self-similarity in real datasets, the above relationship holds for real datasets too. Our experiments on diverse region datasets as well as previous studies [10, 14] confirm that the law strongly holds for lakes, archipelagoes, vegetative ecosystems, urban areas and many others, as shown in Table 4.

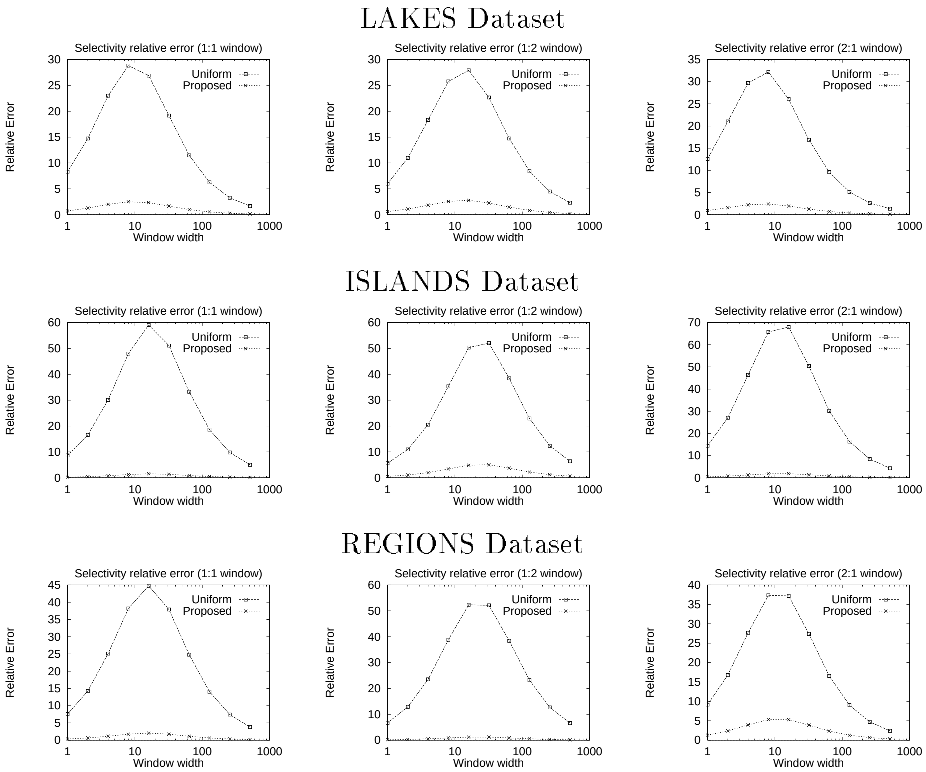


Figure 2: LAKES, ISLANDS and REGIONS (from top to down) datasets: percent relative error vs query window width, for square, 1:2 and 2:1 window queries (from left to right). Proposed method ("$\times$") and uniformity model ("$\Box$").

In conclusion, the theory of σ-fractals is extremely suitable for the study of real regions datasets. The reasons are the following:

1. It leads to regions that obey the power law (4) with some patchiness exponent. As we illustrated, several diverse real datasets obey this law.

2. It provides an easy, recursive algorithm to generate self-similar, realistic region datasets. All we have to do is to choose some values of N_r and r, such that $\log N_r / \log r = B/d$, choose an iterator with that N_r and r, and apply it recursively as many times as needed.

3. It provides useful theorems (like for instance, Theorem 2) which link the patchiness exponent B with the fractal dimension D_H of the boundary of a set of regions. This is important, because we can tap the literature of fractals, where the fractal dimension of several datasets is mentioned (e.g., see appendix in [14, 17]) to obtain an accurate estimation of B.

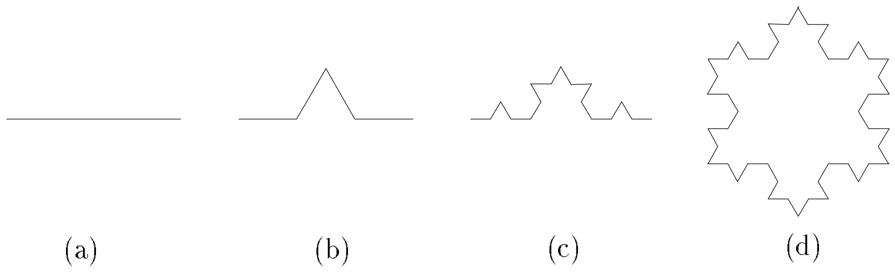


Figure 3: Koch snowflake: (a) initiator; (b) generator; (c) second iteration; (d) relative Koch snowflake.

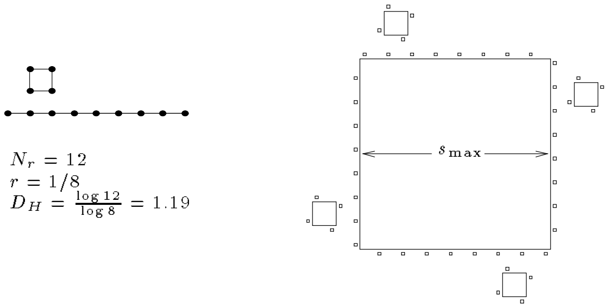


Figure 4: The region generator (left) and the synthetic dataset after two steps of generation (right).

5.2 Directions for a practitioner: fast estimation of B

The final question is: how can a practitioner benefit of our analysis? We have solid answers to this question. In fact, up to now, assuming safely that N and A are given in advance, to estimate selectivity for range queries making use of developed formulas [11, 15], one was required to compute the total width and length extent of the MBRs of the objects. This can be done in two ways: either applying image processing algorithm to the representing bitmap for extracting the regions and compute their MBRs, or, alternatively, by scanning the entire database storing the data. Both approaches are time expensive.

On the contrary, the patchiness exponent B and the ratio ρ can be computed quickly. Concerning ρ, a robust solution is to average the aspect ratios over a small number of regions. Concerning B, we suggest two possible fast ways to compute it, both of them based on sampling. The first makes use of the representing bitmap, while the second works on the database storing the data:

1. Focus on a subwindow of size $t \times t$ of the bitmap, extract the boundaries of the objects contained in it and apply the $O(t \log t)$ time algorithm [1] to compute their fractal dimension D_H. Assuming that regions are self-similar (and then subwindows of the bitmap are similar to the whole) and applying Theorem 2, we can conclude that $B = D_H/2$ is a good approximation for the real B of

Dataset	D_H	B	$D_H - 2B$
LAKES	1.78	0.85	0.08
ISLANDS	1.23	0.60	0.03
REGIONS	1.48	0.70	0.08
Aegean Island	1.08	0.52	0.04
Japan archipelago	1.19	0.59	0.01
Italy plains	1.32	0.63	0.06
Whole Earth [14]	1.2	0.6	0
Cypress vegetation [10]	0.62	1.23	0.01

Table 4: Connection between D_H and B for real datasets: above the line, our own experiments (LAKES, ISLANDS, REGIONS, Aegean Islands, Japan Archipelago and Italy agricultural plains). Below the line, data drawn out from [14, 10], resp.

all the map.

2. Focus on a subwindow of the bitmap, retrieve from the database all the objects contained in it and compute the CCDF of their areas. Then, plot the obtained points in a log-log diagram and interpolate them with a straight line using the classic least-square method. The negated slope of such a line corresponds to the patchiness exponent of the subset of objects. Once again, assuming that regions are self-similar, we can be confident that such exponent is representative for the whole dataset.

Finally, if both approaches are not practicable, we can easily obtain an accurate lower bound on the selectivity by setting $B = 0.5$ and an accurate upper bound by setting $B = 0.9$, since B is experimentally known to range over the interval $[0.5, 0.9]$.

Therefore we conclude that our analysis is suitable in practice and contributes to the solution to the problem of query performance evaluation in real spatial databases.

6 Conclusions

The main contribution of this paper is the accurate modeling of real region datasets, such as archipelagoes, areas of vegetation, city regions, plain maps, hydro-graphic systems and many others. We showed that very few measures are needed (the total count of objects, the total volume, the average aspect ratios among the sides of an object and the patchiness exponent), to achieve extremely accurate results. Our experiments on diverse, real datasets, showed that our approach achieves selectivity estimates within 1-5% for the maximum relative error, against 30-70% of a naive model that uses the uniformity assumption.

We also pinpointed the connection between the patchiness exponent B and fractals, and specifically the σ-fractals. The immediate benefits are (a) a fast method to estimate B and (b) a simple method to generate realistic region data.

Promising future directions include the use of σ-fractals to study selectivities of additional query types (nearest neighbor etc.) and to analyze SAMs on real region data.

Acknowledgments

The authors would like to thank Alice Caraceni for her help in carrying out experiments presented in the paper.

References

[1] A. Belussi and C. Faloutsos. Estimating the selectivity of spatial queries using the 'correlation' fractal dimension. In *Proc. of the 21st VLDB Conference*, pages 299–310, Zurich, Switzerland, 1995.

[2] S. Berchtold, C. Böhm, D.A. Keim, and H.P. Kriegel. A cost model for nearest neighbor search in high-dimensional data space. In *Proc. ACM SIGACT-SIGMOD-SIGART PODS*, pages 78–86, 1997.

[3] S. Chistodoulakis. Implication of certain assumptions in database performance evaluation. *ACM TODS*, June 1984.

[4] D. J. DeWitt, N. Kabra, J. Luo, J. M. Patel, and J.B. Yu. Client-server paradise. In *Proc. of the 20th VLDB Conference*, pages 558–569, Santiago de Chile, Chile, 1994.

[5] C. Faloutsos and I. Kamel. Beyond Uniformity and Independence: Analysis of R-trees Using the Concept of Fractal Dimension. In *Proc. ACM SIGACT-SIGMOD-SIGART PODS*, pages 4–13, Minneapolis, MN, May 1994. Also available as CS-TR-3198, UMIACS-TR-93-130.

[6] C. Faloutsos, M. Ranganathan, and Y. Manolopoulos. Fast subsequence matching in time-series databases. In *Proc. ACM SIGMOD*, pages 419–429, Minneapolis, MN, May 1994. 'Best Paper' award; also available as CS-TR-3190, UMIACS-TR-93-131, ISR TR-93-86 .

[7] C. Faloutsos and S. Roseman. Fractals for secondary key retrieval. In *Eighth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems (PODS)*, pages 247–252, Philadelphia, PA, March 1989. also available as UMIACS-TR-89-47 and CS-TR-2242.

[8] V. Gaede and W.F. Riekert. Spatial access methods and query processing in the object-oriented gis godot. In *Proc. of the AGDM'94 Workshop*, pages 40–52, Delft, The Netherlands, 1994.

[9] R.H. Güting. An introduction to spatial database systems. *VLDB Journal*, 3(4):357–399, 1994.

[10] H.H. Hastings and G. Sugihara. *Fractals.* Oxford Science Publications, 1993.

[11] I. Kamel and C. Faloutsos. On packing R-trees. In *Proc. of the 2nd ACM Intl. Conf. on Information and Knowledge Manage ment*, pages 490–499, Washington, DC, 1993.

[12] J. Korcak. Deux types fondamentaux de distribution statistique. *Bull. de l'Institute International de statistique*, 3:295–299, 1938.

[13] W. Lu and J. Han. Information associated join index for spatial range search. *Int. Journal of Geographical Information Systems*, 9(3):221–249, 1995.

[14] B.B. Mandelbrot. *The fractal geometry of nature.* W.H. Freeman and Company, 1982.

[15] B. Pagel, H. Six, H. Toben, and P. Widmayer. Towards an analysis of range query performances. In *Proc. of the ACM-SIGMOD Symposium on Principles of Database Systems*, pages 214–2210, Washington, DC, 1993.

[16] A. Papadopoulos and Y. Manolopoulos. Performance of nearest neighbor queries in R-trees. In *6th Int. Conf. on Database Theory (ICDT '97)*, Delphi, Greece, January 1997.

[17] M. Schroeder. *Fractals, Chaos, Power Laws: Minutes From an Infinite Paradise.* W.H. Freeman and Company, New York, 1991.

[18] S. Shekhar, M. Coyle, B. Goyal, D. Ren Liu, and S. Sarkar. Data models in geographic information systems. *CACM*, 40(4):103–11, 1997.

[19] V.J. Tsotras, C.S. Jensen, and R.T. Snodgrass. An extensible notation for spatiotemporal index queries. *SIGMOD Record*, 27(1):47–53, 1998.

About the authors

Guido Proietti received his degree of Doctor in Applied Mathematics from the University of Rome 'La Sapienza' in 1990.

In 1996 he joined the Department of Pure and Applied Mathematics at University of L'Aquila, where he currently works as Researcher in Computer Science. In 1998 he has been visiting scientist at the Department of Computer Science of Carnegie Mellon University, Pittsburgh, USA.

His current activity is mainly concerned with the design of efficient spatial access methods. He is also interested in graph algorithms, computational geometry and computer vision.

Christos Faloutsos received is B.Sc. from Nat. Tech. U. Athens and is M.Sc. and Ph.D. in computer science from University of Toronto. In 1998 he joined the Department of Computer Science of Carnegie Mellon University, Pittsburgh, USA, where he currently works as Associate Professor. He served as PC member for several Conferences.

His research interest include: Query by content in multimedia databases, fractals for clustering and spatial access methods, data mining, data base performance evaluation (declustering, buffering etc), medical image databases and searching in text databases.

Contents